Conversión de las tablas del Léxico-Gramática del francés en el léxico *LGLex*

Elsa Tolone
LIGM, Université Paris-Est, Francia & FaMAF, Universidad Nacional de Córdoba, Argentina
elsa.tolone@univ-paris-est.fr

Abstract

En este artículo describimos las tablas del Léxico-Gramática del francés que han sido convertidas en el léxico *LGLex* para poder usarlas en aplicaciones de Procesamiento del Lenguaje Natural (PLN). Presentamos las tablas del Léxico-Gramática de L. Pivaut que reúnen nombres predicativos construidos con el verbo soporte *faire* (*hacer*) que han sido integradas al léxico *LGLex*. También mostramos la conversión al formato Le*fff* que permite integrarlas directamente en el analizador sintáctico FRMG.

1 Introducción

Los volúmenes de información textual disponibles hoy en día hacen imposible el tratamiento manual de la información, con lo cual el tratamiento automático inteligente se convierte en una necesidad. En este artículo queremos describir cómo se ha podido desarrollar un léxico del francés, basándonos en una muestra que nos ha permitido completar este léxico.

El objetivo general es la comprensión del Lenguaje Natural, la mejora de herramientas y recursos básicos para el análisis automático del francés. El impacto está en aplicaciones muy diversas, desde extracción de información a apoyo al aprendizaje de segundas lenguas. Los léxicos sintácticos son recursos básicos en la mayoría de las tareas avanzadas de Procesamiento del Lenguaje Natural (PLN), ya que la mayoría de los sistemas que disponen de cierta capacidad de comprensión del Lenguaje Natural requieren conocimiento sintáctico y semántico a nivel de predicativo (verbo, nombre o adjetivo).

Las tablas del Léxico-Gramática constituyen hoy en día uno de los principales léxicos sintácticos del francés [3]. Estas tablas han sido compiladas originalmente en los años 1970 por M. Gross [1], en el laboratorio LADL (Laboratorio de Automática Documental y Lingüística), y después en el Laboratorio de Informática de Gaspard-Monge (Universidad de Paris-Est) [1, 4] en Francia. Cada tabla corresponde a una *clase* que reúne elementos lexicales de una categoría gramatical (verbos, nombres predicativos, adverbios, expresiones estereotipadas, etc.) que comparten propiedades.

Se han investigado formas para mejorar las tablas del Léxico-Gramática del francés [9], lo que ha permitido construir un léxico del francés, llamado *LGLex* [2]. Después, este léxico ha sido integrado en el analizador sintáctico FRMG [2] [7] convirtiéndolo al formato Le*fff*, el Léxico de las formas flexionadas del francés [6]. El analizador sintáctico FRMG ha sido evaluado con el léxico obtenido en el corpus de referencia de la campaña EASy [10] y Passage [11], usando un módulo integrado a FRMG que permite eliminar la ambigüedad para considerar una sola análisis por frase [3].

[1] El punto de partida de M. Gross es el estudio de las frases simples del francés ya que considera que la unidad mínima de sentido es la frase. El objetivo es repertoriar las frases simples y estudiar las transformaciones que pueden soportar. Las propiedades estudiadas para cada una de esas frases son mayoritariamente propiedades formales sobre la sintaxis más que sobre la semántica, lo que garantiza una reproductividad de tests [3].

[2] Para el anclaje lexical-sintáctico se selecciona un conjunto de árboles de la gramática TAG donde los hypertags de las anclas se unifican correctamente con las palabras del texto de entrada. Se dispone entonces de un conjunto de árboles relacionados con las palabras llenando las funciones principales de la frase de entrada.

[3] Está basado en un algoritmo en programación dinámica de búsqueda de la mejor análisis sumando los pesos de los arcos participando a una análisis. El peso de un arco resulta de la acumulación de los pesos dados por reglas elementales que tienen

En este artículo vamos a presentar una muestra de las tablas de nombres predicativos que hemos añadido al léxico como ejemplo del trabajo realizado para la creación de este léxico. Primero describimos lo que son las tablas del Léxico-Gramática en la sección 2. Después presentamos las tablas del Léxico-Gramática de L. Pivaut en la sección 3 y mostramos cómo las hemos integrado al léxico *LGLex* en la sección 4, explicando el formato y dando un ejemplo del léxico obtenido. Para terminar presentamos el formato L*efff* y un ejemplo del léxico convertido en este formato en la sección 5, antes de concluir en la sección 6.

2 Las tablas del Léxico-Gramática

Una tabla se presenta sobre la forma de una matriz : en línea, los elementos de la clase correspondiente ; en columna, las propiedades sintáctico-semánticas que no son necesariamente aceptadas por todos los miembros de la clase ; al cruce de una línea y de una columna el signo + o − en función de si la entrada léxica descrita por la línea acepta o no la propiedad descrita por la columna. Todos los elementos que figuran en una tabla comparten *propiedades definitorias* de la clase, incluso una *construcción de base*.

Una propiedad sintáctico-semántica es una información que se refiere directamente a la construcción de base asociada a la clase, o una transformación de la construcción de base, o bien una construcción adicional (por ejemplo, las construcciones metafóricas, que no tienen las mismas distribuciones de argumentos). Por ejemplo, la propiedad N0 V significa para un verbo ”posibilidad de ser el predicado de una construcción intransitiva con el sintagma nominal sujeto inicial” , la propiedad [passif par] (por) significa ”diátesis pasiva posible”.

Para ilustrar, la Tab. 1 muestra un extracto de la tabla 33 de los verbos que se construyen con un argumento introducido por la preposición *à* (*a*), lo que significa que reúne los verbos que aceptan la propiedad N0 V à N1.

Si un verbo tiene dos sentidos distintos, tiene dos entradas lexicales ya que cada sentido no acepta el mismo conjunto de propiedades. Uno de los ejemplos que figura en la tabla 33 es el verbo *se rendre* (*rendirse*) :
Max s'est rendu à mon opinion - Max se rindió a mi opinión
Le caporal s'est rendu à l'ennemi - El caporal se rindió al enemigo

Se puede ver que *se rendre* (en el sentido de aceptar) tiene un complemento nominal no humano : la propiedad N1 =: N-hum está aceptada (signo +), mientras no está aceptada (signo −) para *se rendre* (en el sentido de capitular).

Actualmente, para el francés, existen 67 tablas (o clases) de verbos simples, 78 tablas de nombres predicativos simples y compuestos (nombres con argumento(s) que son estudiados con su verbo soporte)[4], 69 tablas de expresiones estereotipadas (principalmente verbales y adjetivales) y 32 tablas de adverbios simples (adverbios en *-ment - -mente*) y (semi-)estereotipados.

El número de entradas es de 13 867 para los verbos (5 738 de los cuales son lemas distintos, si no se tienen en cuenta los diferentes sentidos), 12 696 para los nombres predicativos (8 531 de los cuales son

en cuenta el arco corriente y otros arcos vecinos. Los pesos están elegidos de manera heurística. Por ejemplo, existen reglas que favorecen los arcos que llenan la valencia de un verbo, la presencia de un sujeto delante de su verbo, la inversión del sujeto si algunas condiciones están cumplidas, etc. Otras reglas penalizan las dependencias a larga distancia, las transcategorisaciones innecesarias, algunas construcciones improbables, etc.

[4]En las frases con verbo soporte, no es el verbo que tiene la función predicativa, sino el nombre predicativo (*Luc monte une* **attaque** *contre le fort - Luque monta un* **ataque** *contra el fuerte*), un adjetivo predicativo (*Luc est* **fidèle** *à ses idées - Luque es* **fiel** *a sus ideas*), etc. La distribución del sujeto y eventualmente de sus complementos esenciales dependen de este elemento predicativo.

N0 =: Nhum	N0 =: N-hum	N0 =: Nnr	<ENT>Ppv	Ppv =: se figé	Ppv =: les figé	Ppv =: Neg	<ENT>V	Neg	N0 V	N0 être V-ant	N0 V de N0pc	N1 =: Nhum	N1 =: N-hum	N1 =: le fait Qu P	Ppv =: lui	Ppv =: y	[extrap]	N0idée V Loc N1esprit	<OPT>
+	−	−	*les*	−	+	−	*lâcher Advm*	−	+	−	−	+	−	−	−	−	−	−	*Max les lâche difficilement à Ida*
+	−	−	*<E>*	−	−	−	*renaître*	−	+	+	−	−	+	−	−	+	−	+	*Max renaît au bonheur de vivre*
+	−	−	***se***	+	−	−	***rendre***	−	+	−	−	+	**+**	+	−	+	+	−	***Max s'est rendu à mon opinion***
+	−	−	***se***	+	−	−	***rendre***	−	+	−	−	+	**−**	−	−	−	−	−	***Le caporal s'est rendu à l'ennemi***
+	−	−	*<E>*	−	−	−	*renoncer*	−	−	−	−	+	+	−	−	+	−	−	*Max renonce à son héritage*
+	+	+	*ne*	−	−	+	*revenir*	+	−	−	−	+	−	−	+	−	−	−	*La tête de Luc ne revient pas à Max*

Table 1: Extracto de la tabla 33 de los verbos

distintos), 39 628 para las expresiones estereotipadas (38 658 de las cuales son distintas) y 10 488 para los adverbios (9 326 de los cuales son distintos). Las propiedades codificadas para todas estas entradas se agrupan en una clase de tablas por categoría. Son en total 551 propiedades para los verbos, 496 para los nombres predicativos, 276 para las expresiones estereotipadas y 159 para los adverbios.

Todas las tablas, al igual que los diferentes léxicos generados y las herramientas que han permitido producirlos, están disponibles bajo una licencia libre LGPL-LR : `http://infolingu.univ-mlv.fr` (Languages Ressources > Lexicon-Grammar > Download). Existen tablas similares para otras lenguas como el italiano, el portugués, el griego moderno y el coreano.

Las tablas han sido modificadas para que sean directamente utilizables en aplicaciones de PLN [8]. En efecto, las propiedades definitorias, que son comunes a todas las entradas de una tabla, tienen la particularidad de no estar codificadas en las tablas pero únicamente descritas en la literatura. Para que las tablas sean utilizables, hace falta explicitar las propiedades de cada una y añadirlas a la *tabla de las clases* de la categoría correspondiente, en la cual figura el conjunto de las propiedades. Además, se deben cambiar de nombre muchas propiedades para que sean coherentes entre ellas.

Después, la herramienta *LGExtract* [2] escrita en *Java* permite generar un léxico sintáctico para el PLN a partir de las tablas del Léxico-Gramática. Utiliza la tabla de las clases y un script de extracción que incluye todas las operaciones relacionadas a cada propiedad que deben ser efectuadas para todas las tablas. El léxico *LGLex* que se obtiene tiene vocación a describir las tablas con los conceptos manipulados por ellas, en un formato que cada uno puede hacer cualquiera utilización informática.

Esto ha permitido proyectar la conversión de este léxico al formato Alexina [6], que es el del formato L*efff*. Este formato puede ser usado directamente en aplicaciones de PLN de alto nivel, incluso en aquellas que requieren un análisis sintáctico profundo.

3 Descripción de las tablas del Léxico-Gramática de L. Pivaut

Las tablas del Léxico-Gramática de L. Pivaut [5] describen la sintaxis de nombres predicativos que se construyen con el verbo soporte *faire* (*hacer*). Los nombres que se han tenido en cuenta se refieren a una actividad musical, deportiva o intelectual. Existen 5 tablas que empiezan por FD (Faire Det N) : FD1, FD2, FD3A, FD3B y FD4.

La construcción de base aceptada por las 5 tablas es : N0 faire Det N (sujeto - verbo soporte *faire* - determinante - nombre predicativo) con N0 =: Nhum, que significa que el sujeto siempre puede ser un sustantivo humano, y Det =: du, faire = pratiquer, que significa que el determinante siempre puede ser

partitivo y que el verbo *faire* (*hacer*) tiene el valor semántico del verbo *pratiquer* (*practicar*) :
Max (fait+pratique) du **sport de combat** - *Max (hace+practica)* **deporte de combate** (FD2)

A nivel sintáctico, las tablas se diferencian por :

- la imposibilidad de aceptar el artículo indefinido como determinante para la tabla FD4 :
  *Max fait (de l'***aérobic**+**un aérobic*[5]*) - Max hace (***aeróbic**+**un aeróbic)* (FD4)
  Se traduce por la no aceptación de las dos propiedades Det =: un, faire = fabriquer (fabricar) y Det =: un, faire = pratiquer (practicar).

- la imposibilidad de aceptar un adverbio temporal durativo (Advtd) cuando está presente el artículo indefinido para las tablas FD2, FD3A y FD3B :
  Max a fait (du **football** *pendant un an+*un football pendant un an) - Max hizo (***fútbol** *durante un año+*un fútbol durante un año)* (FD2)
  La propiedad Det =: un, Advtd no está aceptada por las tablas FD2, FD3A, FD3B y igualmente FD4.

- además del punto anterior, el sustantivo predicativo contiene siempre un determinante numeral (Dnum) seguido por una unidad de distancia para las tablas FD3A y FD3B : la entrada es de la forma <ENT>Dnum <ENT>N (*cinq mille mètres - cinco mil metros*) para la tabla FD3A y <ENT>Dnum <ENT>N <ENT>Nc <ENT>Adjc (*cent mètres nage libre - cien metros de nada libre*) para la tabla FD3B (FD3A y FD3B no representan diferentes construcciones, pero diferencias morfológicas de las entradas).

- el hecho de que *faire* sea sinónimo de *fabriquer* (*fabricar*), *faire* aparezca como una extensión aspectual de *avoir* (*haber*), *avoir* con el artículo indefinido sea substituible a *faire* con el artículo indefinido y Dnum sea siempre posible como determinante para la tabla FD1 :
  Max fait du **piano** - *Max toca el* **piano** / *Max (fait+fabrique+a) (un piano+dix pianos) - Max (hace+fabrica+tiene) (un piano+diez pianos)* (FD1)
  Las propiedades Det =: un, faire = fabriquer y Det =: un-Modif, faire = fabriquer están aceptadas por la tabla FD1, así como las propiedades Det =: un, faire = avoir y Det =: Dnum. Al contrario, las propiedades Det =: un, faire = pratiquer y Det =: un-Modif, faire = pratiquer no lo están.

Esto se puede resumir por el hecho de que si el conjunto de propiedades siguientes está aceptado, es una entrada de la tabla FD1 :
Det =: un, faire = fabriquer / Det =: un-Modif, faire = fabriquer / Det =: un, faire = avoir / Det =: Dnum / Det =: un, Advtd

Si este conjunto de propiedades está aceptado, es una entrada de la tabla FD2, FD3A o FD3B :
Det =: un, faire = pratiquer / Det =: un-Modif, faire = pratiquer

Si ninguna de las propiedades mencionadas está aceptada, es una entrada de la tabla FD4.

4 Integración de las tablas del Léxico-Gramática de L. Pivaut

Primero se han digitalizado las tablas, escaneándolas y aplicando una herramienta de reconocimiento óptico de caracteres (OCR), y se han corregido manualmente los errores aparecidos durante la digitalización, al igual que las faltas de ortografía presentes en la versión original. Después, basándonos en la tesis de L. Pivaut [5], se han modificado manualmente las propiedades de las tablas y se ha definido el conjunto de propiedades para las 5 tablas, incluso las propiedades definitorias de cada tabla. Se ha integrado en la tabla de las clases de los nombres predicativos este conjunto de propiedades y se han

[5]El símbolo * marca una frase inaceptable.

añadido en el script de los nombres predicativos de *LGExtract* las nuevas propiedades. Por fin, se ha podido generar directamente el léxico *LGLex* con *LGExtract*, y también en el formato Le*fff* ya que el script de conversión no requiere ningún cambio.

Los errores corregidos en las tablas FD1, FD2 y FD4 son :

- 249 errores aparecidos durante la digitalización :
  musique **a***ilitaire* → *musique* **m***ilitaire* (*música militare*) ;
  piano **è** *queue* → *piano* **à** *queue* (*piano de cola*) ;
  *graphis***t***e* → *graphis***m***e* (*grafista* → *grafismo*).

- 24 entradas corregidas que son faltas de ortografía presentes en la versión original :
  *droit internatio***nn***al* → *droit internatio***n***al* (*derecho internacional*) ;
  *orgue De***rr***eux* → *orgue De***r***eux* (*órgano Dereux*) ;
  *kusari-***k***ama* → *kusari-***g***ama*.

- 2 entradas suprimidas presentes en la versión original :
  ping-pong : supresión de *Ping-Pong* ya que no se distinguen las mayúsculas de las minúsculas ;
  yoseikan-budo : supresión de *yosukan-budo* ya que esta variante ortográfica no está usada.

- 70 nuevas entradas por desdoblar entradas de la versión original :
  posthorn en inglés : *cor postal*, *cor de postillon* y *cornet postal* en francés (*corneta de posta*, *trompa de postillón*, *corneta de postillón* y *trompa de posta* en español) ;
  lancer du javelot : *lancer de javelot* (*lanzado de jabalina*) con una variación de preposición ;
  hatha-yoga : *hathayoga* sin guión y sin espacio.

A continuación, se muestra en la Fig. 1 la tabla tal como aparece en la tesis de L. Pivaut [5] (que corresponde a la versión 1, en formato papel) y en la Tab. 2 la versión corregida (que corresponde a la versión 3, en formato Excel, a partir de la cual se genera el léxico *LGLex*). Se puede ver que hemos fusionado los *cartuchos* para que solo haya una propiedad por columna. Hemos añadido la primera columna que define el número de la entrada y también la propiedad N0 être Nagent que significa que existe un nombre agente derivado que se construye con el verbo *être* (*ser*) : es positiva cuando por lo menos una de las dos columnas suff- y suff+ no está vacía. La documentación en la Tab. 3 permite explicitar el sentido de cada propiedad que está presente en la versión 1.

En la Tab. 4 se puede ver lo que hemos añadido en la tabla de las clases[6] de los nombres predicativos, que corresponde a todas la propiedades de las 5 tablas de L. Pivaut.

En el script de *LGExtract* hemos añadido estas propiedades, que no existían en las tablas anteriores :

- nuevas categorías posibles que forman parte de las entradas : <ENT>Dnum / <ENT>Adjc ;

- nuevas construcciones : N0 être Nagent / N0 jouer à Det N / N0 jouer de Det N ;

- nuevo determinante posible : Det =: Dnum ;

- determinantes que implican un valor semántico a la frase, que se traduce por el hecho de admitir otro verbo soporte en lugar de *faire* : Det =: du, faire = pratiquer / Det =: un, faire = fabriquer

[6]En una tabla de las clases se invierten las lineas y las columnas y en lugar de poner las entradas, se ponen el nombre de las clases. El signo + (resp., −) significa que todas las entradas de esa clase aceptan (resp., no aceptan) la propiedad. El signo o significa que la propiedad figura en la tabla ya que depende de las entradas. El signo O significa que la propiedad tendría que figurar en la tabla ya que depende de las entradas pero no ha sido todavía codificada.

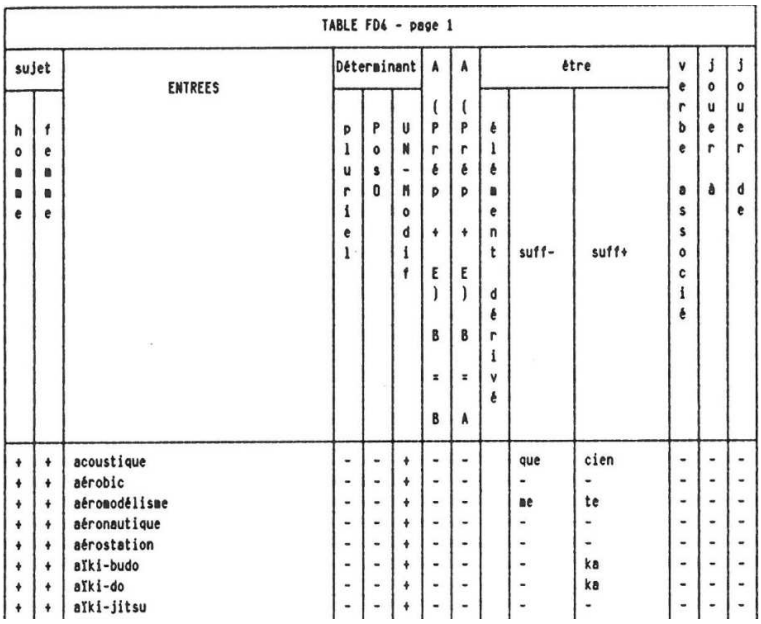

TABLE FD4 - page 1

sujet		ENTREES	Déterminant			A (Prép+E) B = B	A (Prép+E) B = A	être			verbe associé	jouer à	jouer de
homme	femme		pluriel	Poss0	UN-Modif			élément dérivé	suff-	suff+			
+	+	acoustique	-	-	+	-	-		que	cien	-	-	-
+	+	aérobic	-	-	+	-	-		-	-	-	-	-
+	+	aéromodélisme	-	-	+	-	-		me	te	-	-	-
+	+	aéronautique	-	-	+	-	-		-	-	-	-	-
+	+	aérostation	-	-	+	-	-		-	-	-	-	-
+	+	aïki-budo	-	-	+	-	-		-	ka	-	-	-
+	+	aïki-do	-	-	+	-	-		-	ka	-	-	-
+	+	aïki-jitsu	-	-	+	-	-		-	-	-	-	-

Figure 1: Extracto de la tabla FD4 de los nombres predicativos (versión 1)

<ID>	N0 =: homme	N0 =: femme	<ENT>N	Det pluriel obl	Det =: Poss0	Det =: un-Modif, faire = pratiquer	A (Prép+E) B = B	A (Prép+E) B = A	N0 être Nagent	élément dérivé	suff-	suff+	verbe associé	N0 jouer à Det N	N0 jouer de Det N
1	+	+	*acoustique*	−	−	+	−	−	+	−	*que*	*cien*	−	−	−
2	+	+	***aérobic***	−	−	+	−	−	−	−	−	−	−	−	−
3	+	+	*aéromodélisme*	−	−	+	−	−	+	−	*me*	*te*	−	−	−
4	+	+	*aéronautique*	−	−	+	−	−	−	−	−	−	−	−	−
5	+	+	*aérostation*	−	−	+	−	−	−	−	−	−	−	−	−
6	+	+	*aïki-budo*	−	−	+	−	−	+	−	−	*ka*	−	−	−
7	+	+	*aïki-do*	−	−	+	−	−	+	−	−	*ka*	−	−	−
8	+	+	*aïki-jitsu*	−	−	+	−	−	−	−	−	−	−	−	−

Table 2: Extracto de la tabla FD4 de los nombres predicativos (versión 3)

/ Det =: un, faire = pratiquer / Det =: un, faire = avoir / Det =: un-Modif, faire = fabriquer / Det =: un-Modif, faire = pratiquer ;

- obligación a que el determinante sea plural : Det pluriel obl ;
- posibilidad de afinar el sexo del sujeto humano : N0 =: homme / N0 =: femme.

A partir de *LGExtract* se genera el léxico *LGLex* en formato texto o XML [9]. En el formato texto, una entrada se presenta como sigue :

- la entrada comienza por un código que indica su categoría, la tabla donde proviene y el número de la entrada en la tabla (**ID=categoría_númTabla_númEntrada**).

Propiedad	Descripción de la propiedad
N0 =: homme	actividad donde el sujeto practicante es exclusivamente masculino
N0 =: femme	actividad donde el sujeto practicante es exclusivamente femenino
Det pluriel obl	el determinante es obligatoriamente en plural
Det =: Poss0	indica que en paralelo a la frase en *faire en* (*hacer en*) existe una frase en *faire Poss*, el posesivo estando obligatoriamente coreferente al sujeto
Det =: un-Modif	indica la posibilidad del determinante *un-Modif* (*uno* con un modificador)
Det =: un, faire = fabriquer	indica que el determinante *un* (*uno*) es posible, dando a la frase el sentido de *fabriquer* (*fabricar*) además del de *pratiquer ponctuellement* (*practicar puntualmente*)
A (Prép+E) B = B	indica para los nombres compuestos construidos con dos componentes (A,B) que el primero componente (A) puede borrarse
A (Prép+E) B = A	indica para los nombres compuestos construidos con dos componentes (A,B) que el segundo componente (B) puede borrarse
élément dérivé	indica para los nombres compuestos sobre qué elemento se aplica la derivación
suff-	sufijo quitado al nombre predicativo para formar el Nagent (Nagente) derivado
suff+	sufijo añadido al nombre predicativo para formar el Nagent (Nagente) derivado
verbe associé	indica que al nombre predicativo está asociado un verbo
N0 jouer à Det N	indica que en paralelo a la frase *faire de "activité"* (*hacer de "actividad"*) existe una frase *jouer à "activité"* (*jugar a "actividad"*)
N0 jouer de Det N	indica que en paralelo a la frase *faire de "activité"* (*hacer de "actividad"*) existe una frase *jouer de "activité"* (*jugar de "actividad"*)

Table 3: Extracto de la documentación de las propiedades

Propiedad \ tabla	**N_fd1**	**N_fd2**	**N_fd3a**	**N_fd3b**	**N_fd4**
<ENT>Adjc	−	−	−	o	−
<ENT>Dnum	−	−	o	o	−
<ENT>N	o	o	o	o	o
<ENT>Nc	−	−	−	o	−
A (Prép+E) B = A	o	o	−	−	o
A (Prép+E) B = B	o	o	−	−	o
Det =: Dnum	+	O	O	O	−
Det =: du, faire = pratiquer	+	+	+	+	+
Det =: Poss0	−	−	−	−	o
Det =: un-Modif, faire = fabriquer	+	−	−	−	−
Det =: un-Modif, faire = pratiquer	−	+	+	+	o
Det =: un, Advtd	+	−	−	−	−
Det =: un, faire = avoir	+	−	−	−	−
Det =: un, faire = fabriquer	+	o	−	−	−
Det =: un, faire = pratiquer	−	+	+	+	−
Det pluriel obl	o	o	−	−	o
élément dérivé	o	o	−	−	o
N0 =: femme	o	o	o	o	o
N0 =: homme	o	o	o	o	o
N0 =: Nhum	+	+	+	+	+
N0 être Nagent	o	o	−	−	o
N0 faire Det N	+	+	+	+	+
N0 jouer à Det N	o	o	−	−	o
N0 jouer de Det N	o	o	−	−	o
suff-	o	o	−	−	o
suff+	o	o	−	−	o
verbe associé	o	o	−	−	o

Table 4: Extracto de la tabla de las clases de los nombres predicativos

- la sección **lexical-info** indica las informaciones lexicales de la entrada :
  - el lema (que corresponde a la entrada completa, que sea simple o compuesta), y para las entradas compuestas, las diferentes palabras de la entrada con sus categorías gramaticales, así como para algunas entradas nominales, el adjetivo o el verbo morfológicamente derivado del nombre ;
  - también los auxiliares para las entradas verbales, los verbos soportes y los determinantes para las entradas nominales, y las preposiciones (**prepositions** o **locs** para las preposiciones locativas) asociadas a algunos argumentos.
- la sección **args** describe las distribuciones de los diferentes argumentos (sujeto y complementos, repartidos en secciones **const** con la posición **pos**). Una distribución (**comp**) indica :
  - su categoría gramatical : **NP** para un sintagma nominal, **inf** para una infinitiva (V-inf W), **comp** para una completiva (Qu P), **leFaitComp** para el grupo nominal le fait que P (el hecho de que P), **siPOuSiP** para la completiva si P ou si P (si P o si P), **adj** para un adjetivo ;
  - su introductor (**introd-prep** o **introd-loc**) ;
  - rasgos semánticos : **hum** (humano), **nothum** (no humano), **plobl** (plural obligatorio), **npr** (nombre propio), **abst** (abstracto), **conc** (concreto), **source** (origen), **destination** (destinación), **benef** (beneficiario), **mesure** (medida), **prix** (precio), **coll** (colectivo), **plur** (plural).
- la sección **all-constructions** lista las diferentes construcciones aceptadas por la entrada.
- la sección **example** ilustra la entrada (únicamente para los verbos, como se puede ver en la Tab. 1).

Como ejemplo del léxico generado, damos a continuación la entrada *aérobic* (*aeróbic*) que se puede ver en la Tab. 2 y ilustrar por la frase *Max (fait+pratique) de l'aréobic - Max (hace+practica) aeróbic* :

```
ID=N_fd4_2;status=completed
lexical-info=[cat="noun",
              Vsup=[cat="verb",list=(value="faire",
                                     value="pratiquer")],
              noun=[notperm=[complete="aérobic"],noun1="aérobic"]],
              detN=[list-det-modif=(det-modif=[det="du+de l'+de la",modif="false"],
                                    det-modif=[det="un+une",modif="true"]),
              prepositions=()]
args=(const=[pos="0",
             dist=(comp=[cat="NP",hum="true",femme="true",homme="true"])])
all-constructions=[absolute=(construction="true::N0 faire Det N"),
                   relative=(),
                   verbales=(),
                   reductionsGN=()]
example=[example=]
```

Proviniendo de las 5 tablas de L. Pivaut, son 1 742 nuevas entradas (1 674 antes de la corrección manual de la versión original) que se añaden al léxico *LGLex*, lo que hace un léxico de nombres predicativos de 14 438 entradas en total que vienen de 83 tablas.

5 Conversión al formato Le*fff*

El Le*fff* (*Lexique des formes fléchies du français* - Léxico de las formas flexionadas del francés) [6] es un léxico sintáctico de amplia cobertura para el francés y libre bajo licencia LGPL-LR : `http://gforge.inria.fr/projects/alexina/`. Está basado en la arquitectura Alexina de adquisición y modelización de léxicos morfológicos y sintácticos. La herramienta está basada en dos niveles de representación :

- Un *nivel intensional* que factoriza la información léxica, de modo que a cada lema se le asocia una clase morfológica e informaciones sintácticas detalladas.
- Un *nivel extensional*, que se genera automáticamente compilando el léxico intensional, en el que se asocia cada forma flexionada con toda su información morfológica y sintáctica : etiqueta morfológica o el marco de subcategorización de su correspondiente reestructuración, etc.

Una entrada intensional recoge la siguiente información :

- Una *clase morfológica*, la cual define los patrones que construyen todas las formas flexionadas del lema.
- Una *categoría léxica*, que puede ser abierta (adjetivos, adverbios, verbos o nombres, etc.) o cerrada (preposiciones, pronombres o conjunciones, etc.).
- Un *marco de subcategorización*, que muestra explícitamente cómo el lema puede ser utilizado en una determinada construcción sintáctica. Éste lista las funciones sintácticas [7] de los posibles argumentos del lema, y la posible realización de cada una de esas funciones [8].
- Posibles *reestructuraciones*, que definen cómo los marcos de subcategorización profunda se transforman para construir marcos de subcategorización de sintaxis superficial [9].

Como ejemplo damos la entrada intensional siguiente del lema verbal *clarifier* (*clarificar*), que es transitivo directo (dos argumentos con las funciones *Suj* y *Obj*[10]) :

```
clarifier___1      v-er:std
                   100;Lemma;v;
                   <Suj:cln|scompl|sinf|sn,Obj:(cla|scompl|sn)>
                   cat=v;
                   %ppp_employé_comme_adj,%actif,%se_moyen_impersonnel,
                   %passif_impersonnel,%passif
```

Se han convertido las tablas de los verbos simples y de los nombres predicativos al formato L*efff* a partir del léxico *LGLex* [9]. La conversión se hace, mediante el script `lglex2ilex` escrito en *Perl*, en cuatro etapas : identificación de la construcción de base y de sus variantes, construcción de los marcos de subcategorización al formato L*efff*, construcción de listas de reestructuraciones asociadas a cada entrada, informaciones complementarias a añadir a las entradas.

La conversión automática en el formato L*efff* del ejemplo anterior produce esta entrada [11] :

```
aérobic___N_fd4_2 inv
                     100;Lemma;cf;
                     <Suj:cln|sn>;
                     cat=nc,@SujNhum;
                     lightverb=faire|pratiquer;
                     %default
```

[7] Las posibles funciones sintácticas usadas en el L*efff* son las siguientes : *Suj* (sujeto), *Obj* (objeto directo), *Objde* (objeto indirecto introducido por la preposición *de*), *Obja* (objeto indirecto introducido por la preposición *à - a*), *Loc* (locativo), *Dloc* (delocativo), *Att* (atributo), *Obl* y *Obl2* (oblicuos).

[8] Las posibles realizaciones son : clíticas, *cln*, *cla* y *cld* para los casos nominativo, acusativo y dativo; directas, *sn*, *sinf*, *scompl*, *sa* y *qcompl* para los sintagmas nominal, infinitivo, completivo, adjetival y preguntas indirectas; y preposicionales, se construyen de la forma *prep-real*, *prep* siendo una preposición y *real* una realización directa (por ejemplo, *avec-sn - con-sn*).

[9] Las reestructuraciones habituales son *%actif* (para la voz activa), *%passif* (para la voz pasiva) y *%ppp_employé_comme_adj* (cuando el participio funciona como adjetivo) para los verbos, y *%default* para los nombres.

[10] La realización del *Obj* está entre paréntesis, lo que significa que el objeto es facultativo.

[11] La categoría *nc* significa nombre común, *lightverb* contiene los verbos soportes de los nombres predicativos y la *macro* *@SujNhum* contiene el rasgo semántico humano para el sujeto.

Son 2 682 entradas producidas en el formato Le*fff* para estas 5 tablas, lo que suma un total de 31 004 entradas de nombres predicativos en el formato Le*fff* [12].

6 Conclusión

Hemos descrito cómo se han convertido todas las tablas del Léxico-Gramática mostrando el ejemplo de las tablas de L. Pivaut, que todavía se pueden mejorar. Por ejemplo, la propiedad verbe associé indica que al nombre predicativo está asociado un verbo pero no indica cual es el verbo morfológicamente derivado del nombre, información que figura en otras tablas. Las propiedades suff- y suff+ indican el sufijo quitado y añadido al nombre predicativo para formar el Nagent derivado en la construcción N0 être Nagent. De la misma manera, se podría explicitar el nombre derivado completo. También habría que tener en cuenta el hecho de que cuando existen dos variantes para una palabra, con y sin guión, siempre se ha elegido la que lleva el guión, así que habría que generar también la otra, reemplazando el guión por un espacio. Por fin, habría que probar de nuevo el analizador sintáctico FRMG con estas nuevas entradas, para ver si da mejores resultados.

References

[1] Jean-Pierre Boons, Alain Guillet, and Christian Leclère. *La structure des phrases simples en français : Constructions intransitives*. Droz, Genève, Suisse, 1976.

[2] Matthieu Constant and Elsa Tolone. A generic tool to generate a lexicon for NLP from Lexicon-Grammar tables. In Michele De Gioia, editor, *Actes du 27e Colloque international sur le lexique et la grammaire (L'Aquila, 10-13 septembre 2008), Seconde partie*, volume 1 of *Lingue d'Europa e del Mediterraneo, Grammatica comparata*, pages 79–193. Aracne, Rome, Italie, 2010. ISBN 978-88-548-3166-7.

[3] Maurice Gross. *Méthodes en syntaxe : Régimes des constructions complétives*. Hermann, Paris, France, 1975.

[4] Alain Guillet and Christian Leclère. *La structure des phrases simples en français : Les constructions transitives locatives*. Droz, Genève, Suisse, 1992.

[5] Laurent Pivaut. *Verbes supports et vocabulaire technique : sport, musique et activités intelectuelles*. PhD thesis, LADL, Université Paris 7, France, 1989.

[6] Benoît Sagot. The Le*fff*, a freely available and large-coverage morphological and syntactic lexicon for French. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC'10)*, La Valette, Malte, 2010.

[7] François Thomasset and Éric de La Clergerie. Comment obtenir plus des méta-grammaires. In *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN'05)*, Dourdan, France, 2005.

[8] Elsa Tolone. Les tables du Lexique-Grammaire au format TAL. In *Actes de MajecSTIC 2009*, Avignon, France, 2009. (8 pp.).

[9] Elsa Tolone. *Analyse syntaxique à l'aide des tables du Lexique-Grammaire du français*. PhD thesis, LIGM, Université Paris-Est, France, 2011. (340 pp.).

[10] Elsa Tolone and Benoît Sagot. Using Lexicon-Grammar tables for French verbs in a large-coverage parser. In *Proceedings of the 4th Language and Technology Conference (LTC'09)*, pages 200–204, Poznań, Pologne, 2009.

[11] Elsa Tolone, Benoît Sagot, and Éric de La Clergerie. évaluation de lexiques syntaxiques par leur intégration dans l'analyseur syntaxique FRMG. In *Actes du 30ème Colloque Lexique et Grammaire (LGC'11)*, Nicosie, Chypre, 2011. To appear.

[12] El número de entradas es más grande que en el formato *LGLex* porque cada construcción diferente de la construcción de base añade una entrada más en el formato Le*fff* si no es una transformación deductible, como por ejemplo el pasivo.